On the factor ambiguity of MCR problems for blockwise incomplete data sets

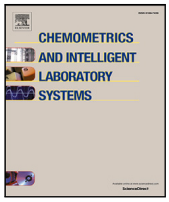
Martina Beese, Tomass Andersons, Mathias Sawall, Cyril Ruckebusch, Gómez Sánchez Adrián, Robert Francke, Adrian Prudlik, Robert Franke, Klaus Neymeyr

Chemometrics and Intelligent Laboratory Systems 249 (2024) 105134

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



On the factor ambiguity of MCR problems for blockwise incomplete data sets

Martina Beese [a,c,*], Tomass Andersons [a], Mathias Sawall [a], Cyril Ruckebusch [g], Adrián Gómez-Sánchez [f,g], Robert Francke [b,c], Adrian Prudlik [b,c], Robert Franke [d,e], Klaus Neymeyr [a,c]

[a] Universität Rostock, Institut für Mathematik, Ulmenstrasse 69, 18057 Rostock, Germany
[b] Universität Rostock, Institut für Chemie, Albert-Einstein-Strasse 3a, 18059 Rostock, Germany
[c] Leibniz-Institut für Katalyse, Albert-Einstein-Strasse 29a, 18059 Rostock, Germany
[d] Evonik Industries AG, Paul-Baumann Strasse 1, 45772 Marl, Germany
[e] Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, 44780 Bochum, Germany
[f] Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028 Barcelona, Spain
[g] Université de Lille, CNRS, Laboratoire de spectroscopie avancé, interactions, réactivité et environnement (LASIRE), F-59000, Lille, France

ARTICLE INFO

Keywords:
Multivariate curve resolution
Area of feasible solutions
Incomplete data
Missing data

ABSTRACT

Multivariate curve resolution (MCR) methods are sometimes faced with missing or erroneous data, e.g., due to sensor saturation. In some cases, an estimation of the missing data is possible, but often MCR works with the largest submatrix without missing entries. This ignores all rows and columns of the data matrix that contain missing values. A successful approach to deal with incomplete data multisets has been proposed by Alier and Tauler (2013), but it does not include a factor ambiguity analysis. Here, the missing data problem is addressed in combination with a factor ambiguity analysis. An approach is presented that minimizes the factor ambiguity by extracting a maximum of spectral information even from incomplete rows and columns of the spectral data matrix. The method requires a high signal-to-noise ratio. Applications are presented for UV/Vis and HSI data.

1. Introduction

When dealing with data sets that contain missing data entries or values that are not suitable for analysis, the question is how to extract the maximum amount of information from the reliable part of the data. For Multivariate Curve Resolution (MCR) problems, maintaining bilinearity during analysis is critical, as it is dealing with different numbers of chemical species. An interesting question is whether and how the spectral information underlying the missing block can be reconstructed. Answers to these questions, as well as a number of applications are presented in this paper for both noise-free and slightly noisy data.

Incomplete data sets or the occurrence of missing values is a common problem that can occur in a wide range of applications, not only in chemistry. For example, blockwise missing data occur if the measurement is outside of the instrument range, the instrument malfunctions for a period of time (e.g., by sensor failure), or a complete measurement is too expensive or difficult for all objects, see [1,2]. This is also the case when multiple measurement techniques are used, for example in image fusion with different resolutions [3] and also for trilinear data from excitation emission measurements, where no emission can be measured below the current excitation. This results in a strong pattern of missing values [4]. Another occurrence is in online process monitoring, where the data of future time steps are unknown and can therefore be assumed to be missing values [2].

Since the goal is to analyze and to extract a maximum of pure component information by MCR methods, a complete spectral data matrix D is required, i.e., there should be no missing values. As stated in [5]: "The most radical – and common – solution is to delete as many variables and/or objects from the data as necessary to reach completeness". However, this results in a loss of information that should be avoided. Since the MCR methods used here are based on a Singular Value Decomposition (SVD), the first approach would be to approximate the SVD of the full matrix.

Several methods have been proposed for this purpose and are widely studied in literature, most are based on statistics. Here, *data imputation* as a concept from statistics is used in combination with iterative algorithms to obtain good approximations of an SVD, e.g., by estimating or using the mean of existing values in the neighborhood of the missing ones, see [2,5–7]. Especially in chemometrics, NIPALS is used to deal with missing values [8,9], but it has limitations with missing data or special patterns. Such patterns can occur when entire blocks of data are missing or when values are not missing at random, e.g., due to an

* Corresponding author at: Universität Rostock, Institut für Mathematik, Ulmenstrasse 69, 18057 Rostock, Germany.
E-mail address: martina.beese@uni-rostock.de (M. Beese).

https://doi.org/10.1016/j.chemolab.2024.105134
Received 18 December 2023; Received in revised form 17 April 2024; Accepted 26 April 2024
Available online 27 April 2024

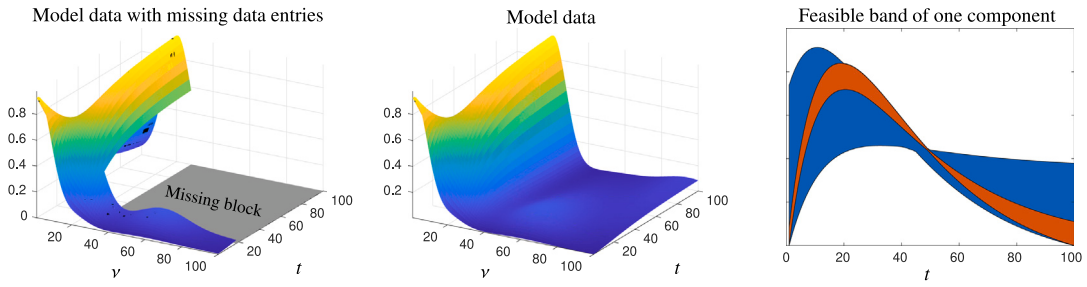


Fig. 1. Simulated data set 1 with missing values (left) and the complete data set (center). Taking only the first 28 frequency channels (the block $D(:, 1:28)$) results in a much wider band of feasible solutions for a certain concentration profile (right, in blue and red color) than using the proposed approach (only the red feasible band). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

upper detection limit. Thus, it is not statistically based and therefore not as common in literature, especially when whole blocks of data are missing rather than sparsely distributed missing entries in the spectral data matrix. Another important aspect is that bilinearity must be maintained to use MCR methods, which can also be a problem of statistical methods. Therefore, other methods that can deal with these properties are needed.

In chemometrics, blockwise missing values and handling bilinearity has been considered by Beyad and Maeder [10]. Their method is based on a linear regression using a given solution of a part of the spectral data matrix. However, such knowledge is not always available. Therefore, Alier and Tauler proposed another approach [11]. Their method obtains a single feasible solution when dealing with incomplete data using a modified version of MCR-ALS where the two factors are calculated iteratively. However, there is a restriction that all chemical components must be shared between the different blocks, see the following subsection for further explanation. This restriction is not needed in the methods presented here. The advantages of the MCR-ALS approach are that it has been tested with different levels of random noise and has been applied in image fusion, e.g., in [3,12], so not only for simulated data sets, but also where noise is present. However, MCR-ALS approximates only one feasible solution and does not represent the full ambiguity.

To obtain a smallest band of feasible solutions, a novel approach is presented using Borgen plots and the Area of Feasible Solutions (AFS), both of which are introduced in Section 1.2.

To the best of our knowledge, no general method for solving this problem using the AFS approach or feasible bands representation has been published. To illustrate the benefit of the new method, see Fig. 1. This simulated data set has more than 50% of missing data, which are concentrated in one block. Looking only at the largest complete subblock, namely a submatrix without missing entries, one gets the range of feasible profiles as shown in blue and red in the right subplot for one concentration profile. However, if we consider all available information, namely the data as shown left in Fig. 1, the ambiguity is drastically reduced (only the red band). This approach is explained in this work.

Our method does not make assumptions or estimates about missing values. Thus, the result is a solution that contains no numerical errors based on the approximation of the SVD. The only requirement for the method is a high signal-to-noise ratio.

A method how these solutions are obtained is explained in Section 2. An application to several data sets follows in Section 3, where a spectroelectrochemical UV/Vis data set, a hyperspectral image data set, and two simulated examples are used to demonstrate the range of applications. First, the mathematical notation of missing blocks and the restrictions on the data matrix are explained, and a brief introduction to the ambiguity of feasible solutions is given.

1.1. Missing blocks in the data matrix

For the explanation of the proposed method, a simple form of missing blocks is assumed, see Eq. (1). This form can easily be extended to more complex structures with more than one missing block, where the matrix can be rearranged or the methods are applied successively, see Section 2.6.

Let $D \in \mathbb{R}^{k \times n}$ be a given matrix for which a submatrix structure is considered. Two different structures are considered here. The first structure is a 2×2 block matrix, where D_{22} is the block of the missing entries. Assuming the existence of a nonnegative matrix factorization $D = CS^T$, namely a pure component factorization, the connecting relations between the various blocks are as follows

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} = CS^T = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}^T = \begin{pmatrix} C_1 S_1^T & C_1 S_2^T \\ C_2 S_1^T & C_2 S_2^T \end{pmatrix} \tag{1}$$

with $D_{11} \in \mathbb{R}^{k_1 \times n_1}$, $D_{12} \in \mathbb{R}^{k_1 \times n_2}$, $D_{21} \in \mathbb{R}^{k_2 \times n_1}$ and $D_{22} \in \mathbb{R}^{k_2 \times n_2}$. The nonnegative factorization couples the pure component information underlying D_{12} and D_{21} to D_{11} via the blocks C_1 and S_1 of the pure component factors, see Eq. (1). Furthermore, the largest submatrices of D without missing values are (D_{11}, D_{12}) and $\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}$. An MCR analysis can be applied to each of these matrices, but they are linked to each other by D_{11}. In the same way, the constraints which restrict an overall solution are connected with D_{11}. Because of this connection and because both matrices share the first block, D_{11} is called the *shared block*. This block is important for the following construction.

A further example to explain the concept of a shared block is a 2×1 block matrix, denoted by

$$D = \begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} S^T \tag{2}$$

with no missing block and $D_{11} \in \mathbb{R}^{k_1 \times n}$ and $D_{21} \in \mathbb{R}^{k_2 \times n}$. Here, both blocks can either be considered separately, or one of them can be used as shared block, since both are coupled by the spectral matrix S. We take D_{11} as shared block. This case was also analyzed in [13], but is included here again as a special case to apply the proposed approach. It is also used as an intermediate step in the construction in Section 2.2. The main difference between our approach and that of [13] is that we do not consider the data set as complete and thus the analysis is not based on the SVD of the complete data set D.

After D_{11} is introduced as shared block, the spectra and concentration profiles of the chemical components underlying D_{11} that are coupled with the other blocks are called shared, in this case *shared chemical components*. Two different cases can be considered, depending on the possible shared chemical components and where they appear.

Case 1: $\operatorname{rank}(D) = \operatorname{rank}(D_{11})$

All chemical components that are present in the full data matrix D are also part of D_{11}.

Case 2: $\operatorname{rank}(D) > \operatorname{rank}(D_{11})$

Not all chemical components appearing in D are shared with D_{11}. For example, D_{11} only depends on a subset of all concentration profiles or spectra.

Note that the individual position of the blocks is not important. When a 2×2 block matrix with one missing block is given, then it can be transformed into the form of Eq. (1). This can be done by using permutation matrices such as $P = \left(\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right)$, where I stands for identity matrices (of potentially different dimensions). Since the following methods focus on the proposed structure, it is recommended to transform a given spectral data matrix into the form of Eq. (1) or (2).

1.2. AFS for the solution ambiguity of MCR problems

The area of feasible solutions (AFS) has a key importance for the analysis of the ambiguity of solutions of the pure component factorization problem in MCR. The following is a brief introduction to this approach, see [14–16] for further details. The goal is to find all feasible factorizations

$$D = CS^T + E, \tag{3}$$

where $D \in \mathbb{R}^{k\times n}$ is a given data matrix, $C \in \mathbb{R}^{k\times s}$ contains the pure component concentration profiles, $S \in \mathbb{R}^{n\times s}$ contains the pure component spectra and $E \in \mathbb{R}^{k\times n}$ describes the error matrix with matrix entries close to zero. In theoretical analyses E is often the null matrix. Otherwise, D is approximated by CS^T. This is the case for numerical calculations or when handling slightly noisy data. The application to experimental data sets with larger errors is planned for future work.

The calculation of the AFS is based on the truncated singular value decomposition (SVD)

$$D = U\Sigma V^T$$

where $U \in \mathbb{R}^{k\times s}$ and $V \in \mathbb{R}^{n\times s}$ each contain the pairwise orthogonal singular vectors, $\Sigma \in \mathbb{R}^{s\times s}$ is a diagonal matrix containing the singular values and s is the number of chemical components.

With a regular matrix $T \in \mathbb{R}^{s\times s}$ the pure concentration and spectral profiles can be reconstructed from the bases of left and right singular vectors

$$C = U\Sigma T^{-1}, \quad S^T = TV^T. \tag{4}$$

However, in most cases there are many feasible matrices T that result in nonnegative C and S and thus many feasible factorizations. This motivates the definition of the AFS for the spectral factor

$$\mathcal{M}_S = \left\{ x \in \mathbb{R}^{s-1}, \text{ so that } W \in \mathbb{R}^{(s-1)\times(s-1)} \text{ exists with } T = \begin{pmatrix} 1 & x^T \\ \mathbf{1} & W \end{pmatrix}, \operatorname{rank}(T) = s \text{ and } C, S \geq 0 \right\}.$$

Thus, the AFS $\mathcal{M}_S$ of the spectral factor is a set in $s-1$ dimensions and represents the so-called rotational ambiguity of the MCR problem, [17–21].

The AFS is bounded by the outer polytope, which represents the nonnegativity constraint (also known as FIRPOL)

$$\mathcal{F}_S = \left\{ x \in \mathbb{R}^{s-1}, \text{ so that } (1, x^T)V^T \geq 0 \right\}$$

and the inner polytope (also known as INNPOL) spanned by the data representing points a_i

$$\mathcal{I}_S = \operatorname{convhull}\left(\left\{ a_i \in \mathbb{R}^{s-1}, \ i = 1, \dots, k \right\} \right), \quad \text{with } a_i = \frac{((U\Sigma)(i, 2:s))^T}{(U\Sigma)(i,1)}. \tag{5}$$

Thus, $\mathcal{F}_S$ and $\mathcal{I}_S$ contain all necessary information about the nonnegativity constraints on C and S that restrict $\mathcal{M}_S$. The a_i are known as *data representing points* or just *data points* in spectral direction, i.e., the i-th row of D is represented by a_i. The data points, the corresponding AFS and the polytopes are illustrated for a typical data set with $s = 3$ chemical species in Fig. 2. This low-dimensional representation is called a Borgen plot, sometimes also called Borgen–Rajkó plot.

The geometric and analytical methods [22,23] only use these polygons to construct the AFS and go back to the work of Borgen and Kowalski [24]. For $s = 3$ chemical species this is a planar construction, which is based on the construction of feasible triangles containing $\mathcal{I}$ and contained in $\mathcal{F}$. Each feasible triangle represents the rows of a matrix T so that C and S according to Eq. (4) are nonnegative matrices. By rotating such feasible triangles in a certain way, it is possible to construct the AFS [25].

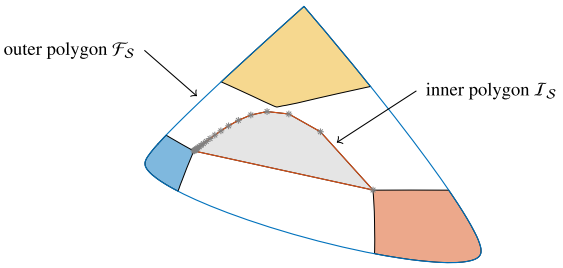


Fig. 2. Low-dimensional representation with the AFS as color-filled areas, the inner and outer polygons (in 2D the polytopes are polygons) as well as the data representing points a_i marked by gray stars. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Several other approaches are based on numerical methods, such as polygon inflation or ray casting [26–28]. They require solving optimization problems to determine whether a point is included in the AFS. In all these cases, the spectral data matrix D is given and is the starting point for all computations.

Once the AFS is known, then the feasible pure component spectra and concentration profiles can be obtained by Eq. (4), where T is built from the vertices of a feasible triangle. Each $x \in \mathcal{M}_S$ represents a feasible spectrum of a pure component, but a valid solution for all components requires a feasible triangle ($s = 3$) or simplex (for $s > 3$).

Similar relations hold for the concentration factor. Here the data points are obtained in frequency direction with $b_j = (V(j, 2:s))^T / V(j, 1)$, which correspond to the columns of D. These define $\mathcal{I}_C$, the dual inner polytope. Also $\mathcal{F}_C$ and $\mathcal{M}_C$ can be defined analogously for the concentration profiles. Both sets are connected via the concept of duality, see [29–32]. For example, if $\mathcal{I}_C$ is given, then $\mathcal{F}_S$ can be uniquely determined.

2. Spectral data matrices with missing blocks

The questions that are answered in this section are: How can one represent a data matrix with missing blocks in a Borgen plot without knowing the entire matrix? How can this be applied to the described cases 1 and 2? How can one get a minimal ambiguity for a given incomplete data set?

Therefore, a shared low-dimensional representation is proposed as well as options for reconstructing the missing data. In order to use a shared low-dimensional representation, it is important to understand essential spectral information and its effect on Borgen plots. Therefore, the concept of *essential spectral information* is first introduced in a general way using convex cones.

2.1. Convex cones and essential spectral information (ESI)

A convex cone C is *finitely generated* if it has the form

$$C = \mathbb{R}_+\{v_1, \dots, v_\ell\} = \left\{ \sum_{i=1}^{\ell} \alpha_i v_i : \ \alpha_i \geq 0 \right\}. \tag{6}$$

Such a finitely generated cone is a polyhedral cone (by the so-called Farkas–Minkowski–Weyl theorem). A cone which is not finitely generated is, for instance, a circular cone for which infinitely many vectors are required for its generation. The expansion in (6) with nonnegative expansion coefficients is called a *conical combination* and presents the cone C based on the vectors $\{v_1, \dots, v_\ell\}$. If the set of

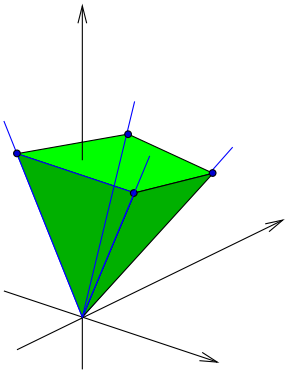


Fig. 3. A polyhedral cone in a 3D space. The edges of the cone are drawn by blue lines and the vertices of the polygonal cross section are marked by blue dots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

generating vectors $\{v_1, \dots, v_\ell\}$ of a polyhedral cone is minimal so that none of the vectors can be omitted without changing C, then the vectors are the edges of the cone. A typical cone with four edges in 3D is shown in Fig. 3.

Let $D \in \mathbb{R}^{k\times n}$ be a nonnegative (spectral data) matrix with the rank s. The cone which is generated by a conical combination of the row vectors of D is given by

$$\text{rowcone}(D) = \left\{ \sum_{i=1}^{k} \alpha_i D(i,:),\ \alpha_i \geq 0 \right\}.$$

Let $\mathcal{E}$ be a smallest subset of indices from $\{1, \dots, k\}$ so that

$$\text{rowcone}(D) = \left\{ \sum_{i\in\mathcal{E}} \alpha_i D(i,:),\ \alpha_i \geq 0 \right\}.$$

The vectors $D(i,:)$ for $i \in \mathcal{E}$ are the edges of $\text{rowcone}(D)$. Further, let $D = U\Sigma V^T$ be an SVD of D. Then the projections of the row vectors of D to the V-space are

$$e_i^T DV = e_i^T U\Sigma.$$

The Perron–Frobenius theory [33] guarantees that the first components of all these vectors are nonzero, under the weak assumption that $D^T D$ is an irreducible matrix, as explained in [27]. This justifies a normalization and to consider only the components $2, \dots, s$, as done in Eq. (5) for the data points a_i. Their convex hull is the inner polygon in the spectral space $\mathcal{I}_S$.

Theorem 2.1. *$D(j,:)$ is an edge of $\text{rowcone}(D)$ if and only if a_j is a vertex of $\mathcal{I}_S$.*

Proof. The following equivalences hold:

$D(j,:)$ is an edge of $\text{rowcone}(D)$

$\Leftrightarrow$ $D(j,:)$ is not representable by the other edges as $\sum_{i\in\mathcal{E},\, i\neq j} \alpha_i D(i,:)$ with nonnegative coefficients α_i

$\Leftrightarrow$ the V-space projection $e_j^T DV = e_j^T U\Sigma$ is not representable as $\sum_{i\in\mathcal{E},\, i\neq j} \alpha_i e_i^T DV = \sum_{i\in\mathcal{E},\, i\neq j} \alpha_i e_i^T U\Sigma$

$\Leftrightarrow$ the data representative a_j is not representable as $\sum_{i\in\mathcal{E},\, i\neq j} \alpha_i a_i$

$\Leftrightarrow$ a_j is a vertex of $\mathcal{I}_S$. □

Hence, considering edges of $\text{rowcone}(D)$ in the high-dimensional space $\mathbb{R}^n$ is one-to-one related to considering vertices of $\mathcal{I}_S$ in the low-dimensional space $\mathbb{R}^{s-1}$. The properties can also be considered with respect to the column space of D in a sense that $D(:,j)$ is an edge of the cone $\text{colcone}(D)$ generated by the column space of D if and only if b_j is a vertex of $\mathcal{I}_C$. The related set of indices is denoted by $\tilde{\mathcal{E}}$.

With these sets of indices, we can define the so-called essential spectral information [34,35]. The formal definition of essential spectra and frequency channels in terms of the cone notation is as follows.

Definition 2.2. A spectrum is called essential, if it is an edge of $\text{rowcone}(D)$, respectively a vertex of $\mathcal{I}_S$.

A frequency channel is called essential, if it is an edge of $\text{colcone}(D)$, respectively a vertex of $\mathcal{I}_C$.

Therefore $\mathcal{E}$ and $\tilde{\mathcal{E}}$ denote the essential spectra and frequency channels. The Cartesian product $\mathcal{E}\times\tilde{\mathcal{E}}$ is denoted by esi, the essential spectral information (ESI) of a spectral data matrix. For a graphical representation see Fig. 2 in [34]. This also means that only the entries of the data matrix corresponding to an essential row as well as an essential column are needed to represent the ambiguity of the matrix factorization. The essential rows are the essential spectra and the essential columns denote the frequency channels. This means that if a matrix is expanded and the new rows or columns have no impact on $\mathcal{I}_S$ respectively $\mathcal{I}_C$, then not only the ESI remain the same, but also the set of solutions remains the same. There is no gain of information. For this reason it is sufficient to consider only the ESI when looking at the ambiguity of the solutions.

Instead of calculating the ESI for an entire matrix, it is also possible to divide the matrix into blocks and determine the ESI for them to simplify the calculation, e.g., if the resulting systems have fewer chemical components than the entire data matrix. The following theorem describes this.

Theorem 2.3. *Let $D = \begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}$ be given. Then the sets of essential spectra $\mathcal{E}_{D_{11}}$, $\mathcal{E}_{D_{21}}$ and $\mathcal{E}_D$ satisfy*

$$\mathcal{E}_D \subseteq \mathcal{E}_{D_{11}} \cup \mathcal{E}_{D_{21}}.$$

Proof. As in Eq. (2) $D \in \mathbb{R}^{k\times n}$ and $D_{11} \in \mathbb{R}^{k_1\times n}$, $D_{21} \in \mathbb{R}^{k_2\times n}$, where $k = k_1 + k_2$.

$$\begin{aligned}
\mathcal{E}_D =& \Big\{ i \in \{1,\dots,k\} \text{ so that } D(i,:) \text{ is not a conical combination} \\
& \quad \text{of the edges of } \text{rowcone}(D) \text{ without } D(i,:) \Big\} \\
\subseteq& \Big\{ i \in \{1,\dots,k_1\} \text{ so that } D(i,:) \text{ is not a conical combination} \\
& \quad \text{of the edges of } \text{rowcone}(D_{11}) \text{ without } D(i,:) \Big\} \\
& \cup \Big\{ i \in \{k_1+1,\dots,k\} \text{ so that } D(i,:) \text{ is not a conical combination} \\
& \quad \text{of the edges of } \text{rowcone}(D_{21}) \text{ without } D(i,:) \Big\} \\
=& \mathcal{E}_{D_{11}} \cup \mathcal{E}_{D_{21}} \quad \square
\end{aligned}$$

The statement is equivalent for the essential frequency channels using D^T. Another way to describe this is to look directly at the complete block matrix. Since merging data does not create new data, i.e., new information, it can still be represented by the essential spectra, respectively frequency channels, of the individual data sets. In the context of missing values, this theorem means, that the reduced versions of the submatrices using $\mathcal{E}$, respectively $\tilde{\mathcal{E}}$, are required for further factor ambiguity calculations.

Remark 2.4. Theorem 2.3 can be extended to any number of submatrices that are used to build a matrix D and its ESI.

A special and important interpretation of the block matrix case is the row-wise extension of D. We are not only interested in the ESI of the extended matrix, but also how the ESI changes when adding rows. Therefore, we first look at how the essential frequency channels are affected and thus at the properties of $\text{colcone}(D)$. We show that the edges of $\text{colcone}(D)$ relate in a one-to-one manner to the edges of the column-space cone $\text{colcone}(S^T)$ assuming the factorization $D = CS^T$ to be given.

Theorem 2.5. *Let $D = CS^T$ be a nonnegative rank-factorization of $D \in \mathbb{R}^{k\times n}$ with $\operatorname{rank}(D) = s$ and $C \in \mathbb{R}^{k\times s}$ and $S \in \mathbb{R}^{n\times s}$. Then $D(:,\ell)$ is an edge of $\operatorname{colcone}(D)$ if and only if $S^T(:,\ell)$ is an edge of $\operatorname{colcone}(S^T)$.*

Proof. We show the complementary case of a certain column not being an edge. It holds that:

$$
\begin{aligned}
& D(:,\ell) \text{ is not an edge of } \operatorname{colcone}(D) \text{ as a conic combination exists,} \\
& D(:,\ell) = (CS^T)e_\ell = \sum_{i\in\tilde{\mathcal{E}}} \alpha_i CS^T e_i,\ \alpha_i \geq 0 \\
\Leftrightarrow\ & C\left(S^T e_\ell - \sum_{i\in\tilde{\mathcal{E}}} \alpha_i S^T e_i\right) = 0,\ \alpha_i \geq 0 \\
\Leftrightarrow\ & S^T e_\ell - \sum_{i\in\tilde{\mathcal{E}}} \alpha_i S^T e_i = 0,\ \alpha_i \geq 0 \text{ (since the } s \text{ columns of } C \text{ are} \\
& \quad \text{linearly independent)} \\
\Leftrightarrow\ & \text{the } \ell\text{th column of } S^T \text{ is a conic combination of the edges} \\
& \quad \text{of } \operatorname{colcone}(S^T). \ \square
\end{aligned}
$$

Next we show that adding further rows to D (e.g., further measured spectra extend the spectral data matrix) does not change the edges of $\operatorname{colcone}(S^T)$, if the rank of D is preserved under the row augmentation.

Theorem 2.6. *Let $D \in \mathbb{R}^{k\times n}$ be a matrix of rank s having a nonnegative factorization $D = CS^T$ with $C \in \mathbb{R}^{k\times s}$ and $S \in \mathbb{R}^{n\times s}$. A $(k+1)$th row is added to D resulting in $\tilde{D} \in \mathbb{R}^{k+1\times n}$, where $\operatorname{rank}(D) = \operatorname{rank}(\tilde{D}) = s$ is assumed. Presuming this added row is also a nonnegative linear combination of the rows of S^T and thus $\tilde{D}$ has also a nonnegative factorization $\tilde{D} = \tilde{C}S^T$ with the same factor S^T.*

Then the two ESI index sets of $\operatorname{colcone}(D)$ and $\operatorname{colcone}(\tilde{D})$ are the same. Further, it holds that $\mathcal{I}_C(D) = \mathcal{I}_C(\tilde{D})$ (and their vertices are the same), which is equivalent to $\mathcal{F}_S(D) = \mathcal{F}_S(\tilde{D})$.

Proof. Applying Theorem 2.5 to $D = CS^T$ and to $\tilde{D} = \tilde{C}S^T$ shows that the edges of $\operatorname{colcone}(S^T)$ are equal to the edges of $\operatorname{colcone}(D)$ and also equal to the edges of $\operatorname{colcone}(\tilde{D})$. The statement for the inner polygons $\mathcal{I}_C(D)$ and $\mathcal{I}_C(\tilde{D})$ follows from Theorem 2.1. Further $\mathcal{F}_S(D) = \mathcal{F}_S(\tilde{D})$ follows from their duality to $\mathcal{I}_C(D)$ and $\mathcal{I}_C(\tilde{D})$. $\square$

The argumentation of this theorem is analogous for adding columns, using D^T. It can also be rephrased for ESI, where the essential frequency channels of D and $\tilde{D}$ are the same. Further information regarding a visualization of the impact on the low-dimensional representation is provided in Section 2.3.

2.2. Representation of the factor ambiguity in terms of convex cones

The next step is to look beyond the inner and outer polygons and to inspect the resulting feasible solutions and how they are affected. Therefore, some additional connections are made next. Here $\operatorname{colcone}(D(1:k,:))$ denotes the cone of the column space of the first k rows of D, and $(\operatorname{colcone}(D))(1:k)$ is a k-dimensional representation of $\operatorname{colcone}(D)$, where only the first k dimensions are considered.

Remark 2.7. Presuming the assumptions of Theorem 2.6 are met, then $(\operatorname{colcone}(D))(1:k) = \operatorname{colcone}(D(1:k,:))$ holds. This follows from the definition of a convex cone when $\tilde{\mathcal{E}}$ is the same (which it is according to Theorem 2.6).

So it is sufficient to look at the first k rows of D when one is only interested in the first k rows of C. But in order to see if C and the corresponding S are solutions, i.e., a nonnegative factorization, we first clarify what is a feasible solution in the context of convex cones.

Theorem 2.8. *If $C \in \mathbb{R}^{k\times s}$ and $D \in \mathbb{R}^{k\times n}$ are nonnegative matrices with $\operatorname{colcone}(C) \supseteq \operatorname{colcone}(D)$, then a nonnegative matrix $S \in \mathbb{R}^{n\times s}$ exists so that $D = CS^T$ and therefore C and S are feasible solutions.*

Proof. $\operatorname{colcone}(C) \supseteq \operatorname{colcone}(D)$ means that every column of D can be represented by a conical combination of the columns of C. These nonnegative linear factors can be stored in a matrix S so that $D = CS^T$. This also means that S is a nonnegative matrix, thus C and S form a nonnegative factorization of D and are therefore feasible solutions. $\square$

This is equivalent to the existence of a nonnegative $S \in \mathbb{R}^{n\times s}$ with $\operatorname{rowcone}(S^T) \supseteq \operatorname{rowcone}(D)$, which describes the transposed case. This can be interpreted in a way that the s edges of $\operatorname{rowcone}(S^T)$ and $\operatorname{colcone}(C)$ enclose the $\operatorname{rowcone}(D)$ and $\operatorname{colcone}(D)$ respectively.

The next theorem describes the impact of additional data, i.e., added rows or columns, on the set of feasible solutions of D.

Theorem 2.9.

Let $D \in \mathbb{R}^{k\times n}$ be a rank-s matrix having a nonnegative factorization. By adding a $(k+1)$th row to D the matrix $\tilde{D} \in \mathbb{R}^{k+1\times n}$ is formed. We assume that $\operatorname{rank}(D) = \operatorname{rank}(\tilde{D}) = s$ and $\tilde{D} = \tilde{C}S^T$ with $\tilde{C} \in \mathbb{R}^{(k+1)\times s}$ and $S \in \mathbb{R}^{n\times s}$.

Then the feasible solutions of $\tilde{D}$ are the same if $\operatorname{colcone}(D)$ and $\operatorname{rowcone}(\tilde{D})$ are considered instead of $\operatorname{colcone}(\tilde{D})$.

Proof. Remark 2.7 shows that $\operatorname{colcone}(D) = \operatorname{colcone}(\tilde{D})(1:k)$ and thus $\operatorname{colcone}(\tilde{C})(1:k) \supseteq \operatorname{colcone}(\tilde{D})(1:k)$. That is, the set of feasible solutions for the first k rows of $\tilde{C}$ is the same for $\operatorname{colcone}(\tilde{D})$ and $\operatorname{colcone}(D)$. Thus, $\tilde{C}(1:k,:)$ and S are the same for both $\operatorname{colcone}(D)$ and $\operatorname{colcone}(\tilde{D})$ because of Theorem 2.8. It remains to be shown that $\tilde{C}(k+1,:)$ is also the same whether $\operatorname{colcone}(D)$ or $\operatorname{colcone}(\tilde{D})$ is considered. However, $\operatorname{colcone}(\tilde{D})$ is not determined by $\operatorname{colcone}(D)$. But it can be reconstructed uniquely by using $\tilde{D}(k+1,:) = \tilde{C}(k+1,:)S^T$. Thus, the resulting feasible solutions are the same and it is sufficient to consider $\operatorname{colcone}(D)$ instead of $\operatorname{colcone}(\tilde{D})$. $\square$

Therefore, it is also possible to consider the original colcone and only change the rowcone (add new edges) to calculate the set of nonnegative factorizations of the extended data matrix. If D is extended by more than one row the proof is analogous. Thus this theorem can be used to represent the solutions of $\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}$ using $\operatorname{colcone}(D_{11})$ and $\operatorname{rowcone}\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}$. The same can be done for D_{12} using $\operatorname{rowcone}(D_{11})$ and $\operatorname{colcone}(D_{11}, D_{12})$. With this it is possible to express the set of nonnegative factorizations of D in an alternative way as shown in the following theorem.

Theorem 2.10. *Under the assumption that the shared block $D_{11} \in \mathbb{R}^{k_1\times n_1}$ has the same rank as $D \in \mathbb{R}^{k\times n}$ and D has a nonnegative factorization, the following holds:*

$$
\begin{aligned}
&\operatorname{colcone}(D)(1:k_1) = \operatorname{colcone}(D_{11}, D_{12}) \\
&\operatorname{rowcone}(D)(1:n_1) = \operatorname{rowcone}\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}
\end{aligned}
$$

Thus the set of nonnegative factorizations $D = CS^T$ is determined by these cones.

Proof. Applying Theorem 2.9 in row direction shows that $\operatorname{colcone}(D)(1:k_1) = \operatorname{colcone}(D_{11}, D_{12})$ holds and the rowcone stays the same. The same can be done in column direction where the rowcone becomes $\operatorname{rowcone}(D)(1:n_1) = \operatorname{rowcone}\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}$. Since this step does not impact the colcone, the desired representation has been achieved. The sets of solution are also the same because of the former theorem and since the factorization of D is uniquely determined with the resulting C_1 and S_1. This applies since the factors C_2 and S_2 are uniquely determined with $D_{12} = C_1 S_2^T$ and $D_{21} = C_2 S_1^T$ and therefore also the complete factorization of D. $\square$

We conclude that in order to calculate all feasible solutions $C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$ and $S^T = (S_1^T, S_2^T)$ the entire matrix D is not needed. This allows to handle cases as described in Eq. (1).

Since it is not very descriptive to represent the solutions of the data set in terms of convex cones we go back to the well known low-dimensional representation in the AFS space and transfer the properties gained in this section.

2.3. Representation of the factor ambiguity in the AFS space

The advantage of using a low-dimensional representation, such as a Borgen plot for $s = 3$, is that the feasible solutions can be represented and a possible solution can easily be selected. The disadvantage is that if the matrix D is changed, e.g., by adding a new row, the SVD changes and so does the low-dimensional representation.

However, the convex cone theory opens a way to preserve the low-dimensional representation. For simplicity, we will only consider the case $s = 3$ in this section. However, the results are also applicable to $s \neq 3$.

Therefore, the results of Sections 2.1 and 2.2 can be transformed to the low-dimensional representation as follows. Here $\mathcal{F}$ and $\mathcal{I}$ always denote the polygons of D_{11}.

Remark 2.11. The statement of Theorem 2.10 can be reformulated for Borgen plots, meaning that it is sufficient to expand $\mathcal{I}_S$ and $\mathcal{I}_C$ by the corresponding data points of D_{21} and D_{12} respectively from which the new outer polygons can be calculated by using duality. Then the AFS can be calculated with these expanded polygons.

This already includes the statement of Theorem 2.9, which is equivalent to:

When expanding D by a new row as in Theorem 2.9, $\mathcal{F}_S$ stays the same and $\mathcal{I}_S$ is given by the convex hull of the data point corresponding to the new row and the old $\mathcal{I}_S$. Then the AFS can be calculated by using the expanded polygons. This is illustrated in Fig. 4 in the upper row.

The data points corresponding to new columns and rows, which means the addition of D_{12} and D_{21} to D_{11}, can be calculated as explained in the following remark.

Remark 2.12. From the SVD $D = U\Sigma V^T$, the data points spanning $\mathcal{I}_S$ can be calculated by $U\Sigma = DV$ followed by scaling the contribution of the first singular vector to the value 1. Thus, the data point corresponding to an extended row x^T can be calculated with $x^T V$ followed by scaling. When additional columns are considered, the same holds for using $V^T = \Sigma^{-1}U^T D$. Thus, any spectra or frequency channel that do not increase the rank can be displayed in the Borgen plot of the original matrix D. This process can be repeated if more than one row or column is added.

This means that it is possible to represent the additional information underlying D_{21} and D_{12} in the same low-dimensional space as defined for D_{11} with full information preservation. This is shown in Fig. 4 where at first D_{21} and then D_{12} are added in order to show their individual and combined impact. The associated feasible bands are shown in Fig. 5, where the feasible bands of the spectral factor are visualized in the same order as in Fig. 4. Since for D_{11} and (D_{11}, D_{12}) the AFS consists of one connected set and not three separated sets, there is only one set of feasible bands. As the feasible bands get smaller when more information is included, the ambiguity decreases. This process starts with the AFS of D_{11} and then the information from D_{12} and D_{21} is added. See the proof of Theorem 2.10 for details. We call this form of representation the *shared low-dimensional representation*, since all information is represented in the low-dimensional representation of the shared block D_{11}. However, not every addition leads to reduced ambiguity. Only when D_{12} and D_{21} contain additional ESI does their inclusion reduce the AFS.

It is possible to select any rank-s submatrix of D and to use it as D_{11} for representing the complete data matrix in the corresponding low-dimensional representation. Thus even a full-rank $s \times s$ submatrix is sufficient to include all the information of the complete matrix. The remaining rows and columns then correspond to D_{12} and D_{21} after a permutation to gain the form of Eq. (1). If this is done in a way that D_{22} contains missing data, then it is now possible to represent a matrix with missing entries in a low-dimensional representation and to calculate the feasible solutions.

The approach can also be applied if the signal-to-noise ratio is not too small. In cases when certain blocks have a low signal-to-noise ratio, it may be beneficial to disregard them, reducing the problem from Eq. (1) to Eq. (2). An application to noisy data sets is provided in Section 3.

The detailed explanation of how to apply this theory to the two cases presented in Section 1.1 regarding the rank of D_{11} follows next.

2.4. Case 1: All chemical species contribute to the shared block D_{11} or $\operatorname{rank}(D) = \operatorname{rank}(D_{11})$

Using the shared low-dimensional representation described above, it is now easy to get a representation of a data set for case 1 where $\operatorname{rank}(D_{11}) = \operatorname{rank}(D)$ and to compute the AFS. However, the restrictive assumption $\operatorname{rank}(D_{11}) = \operatorname{rank}(D)$ allows us to reconstruct a representative of the missing block D_{22} in a unique way. There may be different matrices D_{22}, but all of them have the same pure component factors. Then the factor ambiguity problem can be treated in a usual way, e.g., by using FACPACK, see [27].

Theorem 2.13. *Let* $\operatorname{rank}(D_{11}) = \operatorname{rank}(D)$ *be given with* $D_{11} = C_1 S_1^T$. *Then the missing block* D_{22} *is uniquely determined by* D_{11}, D_{12} *and* D_{21}.

Proof. 1. According to Eq. (1) $D_{12} = C_1 S_2^T$ holds. Thus S_2^T is given by

$$S_2^T = C_1^+ D_{12} \tag{7}$$

with the pseudo-inverse C_1^+ of C_1. Since $\operatorname{rank}(D) = \operatorname{rank}(D_{11})$, the column space of D_{11} includes the column space of D_{12}. Thus the matrix equation $D_{12} = C_1 S_2^T$ can be solved and has the solution (7). The pseudoinverse C_1^+ guarantees a solution of the smallest Euclidean norm.

2. An analogous argumentation allows us to determine C_2 from $D_{21} = C_2 S_1^T$

$$C_2 = D_{21}(S_1^T)^+. \tag{8}$$

The combination of (7) and (8) yields

$$D_{22} = C_2 S_2^T = D_{21}(S_1^T)^+ C_1^+ D_{12} = D_{21}(C_1 S_1^T)^+ D_{12} = D_{21} D_{11}^+ D_{12}.$$

The resulting D_{22} is determined in a unique way by solving the two linear systems with the pseudoinverse, which guarantees solutions with the smallest Euclidean norm. However, the solution D_{22} is not necessarily unique, since there are other solutions with contributions from the respective null spaces. □

This reconstruction works for noise-free data. Otherwise, D_{22} may have negative entries. Next we treat the more complex case, when the shared block D_{11} does not have the full rank, namely if not all chemical species contribute to the data in D_{11}

2.5. Case 2: The shared block D_{11} does not represent all chemical species or $\operatorname{rank}(D) > \operatorname{rank}(D_{11})$

If not all chemical components contribute to D_{11}, we can partially trace the problem back to case 1 in a way that we use a shared low-dimensional representation and find its minimal ambiguity. The approach is described in the following theorem. Certain affine subspaces are required, and these are denoted by the letter $\mathcal{H}$.

Theorem 2.14. *Let* $D = CS^T$ *be given with nonnegative factors* C *and* S *and* $\operatorname{rank}(D) > \operatorname{rank}(D_{11})$. *Then the constraints on the chemical species that are shared between either* D_{11} *and* D_{21} *or* D_{11} *and* D_{12} *can be represented by the Borgen plot of* D_{11}, *the shared block.*

Proof. D_{11} and D_{21} are considered first, with $\operatorname{rank}\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix} > \operatorname{rank}(D_{11})$. The case of columnwise extension with D_{12} can be treated in an

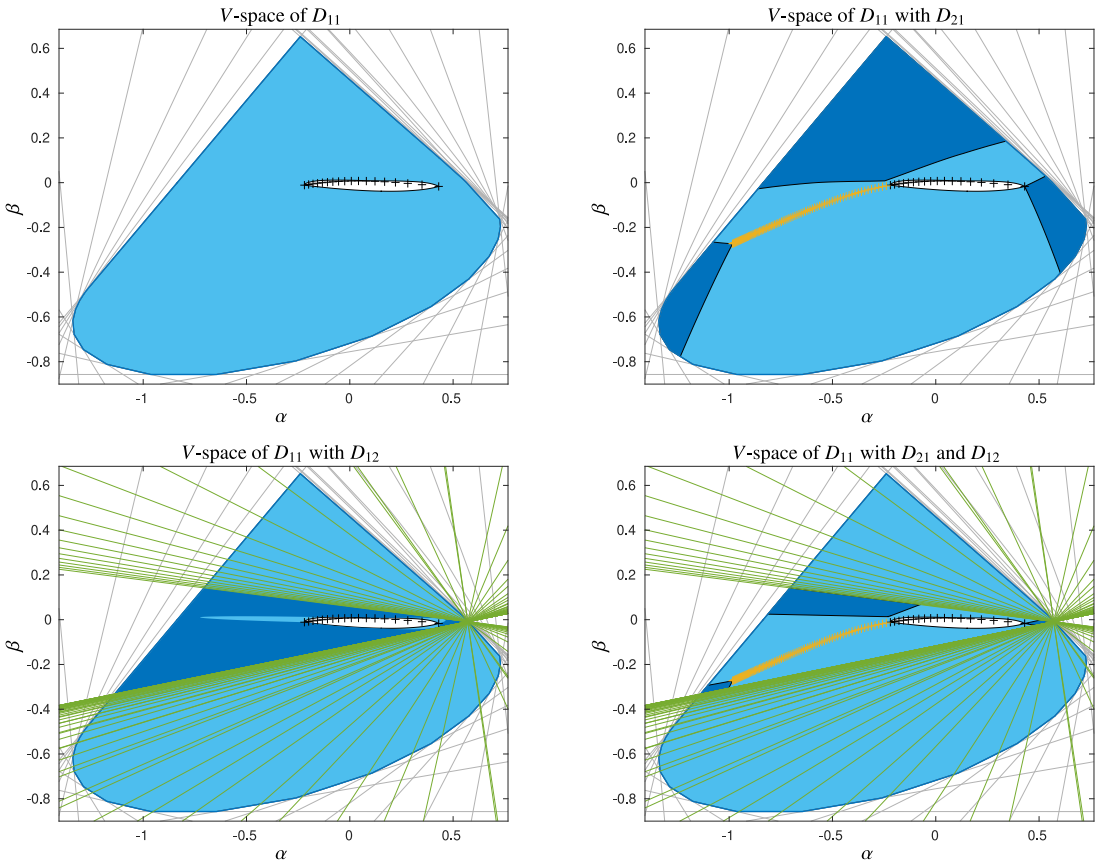


Fig. 4. Shared low-dimensional representations of the simulated data set 1. Top row (left): Low-dimensional representation of D_{11} where the AFS is the large connected set in light blue (with a hole around the origin). Top row (right): D_{11} with added information of D_{21} in yellow results in a much smaller AFS which is plotted in dark blue. Bottom row (left): D_{11} with added information of D_{12} in green and the reduced AFS in dark blue. Bottom row (right): Shared low-dimensional representation of D_{11} with added information of D_{21} and D_{12}. The resulting AFS in dark blue has the smallest size and results in a smallest ambiguity of the nonnegative matrix factorization problem. The corresponding feasible bands are shown in Fig. 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

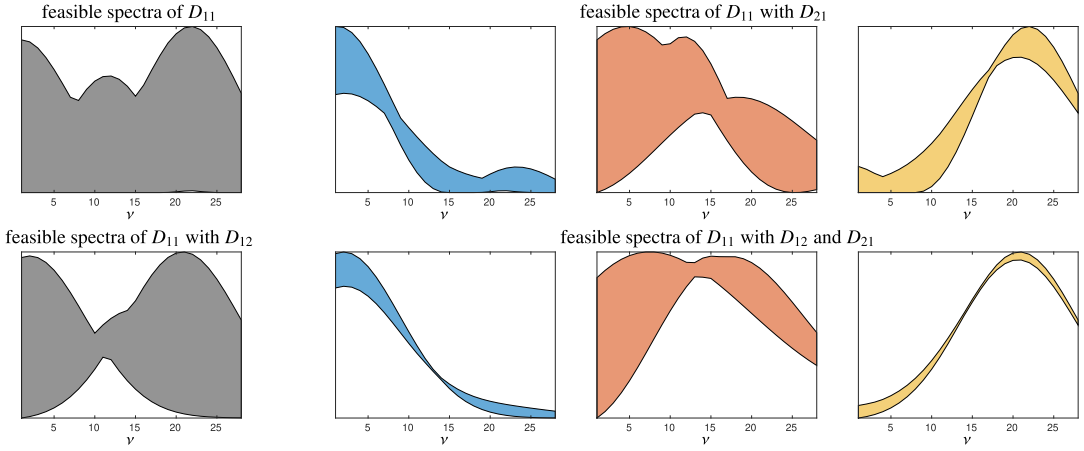


Fig. 5. Normalized feasible spectral bands of the three chemical species represented by the factor S for the feasible regions presented in Fig. 4. The feasible spectral band of D_{11} is shown left in the top row. The AFS is one connected set so that the band is wide. By adding the information of the block D_{21} the AFS decomposes into three isolated subsets, which results in three separate spectral band plots (colored plots in the top row). The three right plots in the bottom row show the feasible bands belonging to the shared low-dimensional representation with the additional information of both D_{21} and D_{12}. This represents the minimal ambiguity that can be achieved. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

analogous way. The starting point is a Borgen plot of $\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}$ in a space with the dimension $\operatorname{rank}\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix} - 1$. (This uses a generalization of the 2D Borgen plots for systems with three chemical species.) Within this (higher-dimensional) space the data points of D_{11} span a $(\operatorname{rank}(D_{11})-1)$-dimensional affine subspace which is denoted by the letter $\mathcal{H}$. Each spectrum in $\mathcal{H}$ can be represented by a linear combination of the spectra underlying the block D_{11}. Those spectra which correspond to data points of D_{21} that are located in the affine subspace $\mathcal{H}$ refer to chemical species that are shared. We collect them in a submatrix D'_{21}, which is a reduced version of D_{21}, with $\operatorname{rank}(D'_{21}) \leq \operatorname{rank}(D_{11})$. With this construction, all constraints which are expressed in terms of $\mathcal{I}$ and $\mathcal{F}$ of D_{11} are included because

1. $\mathcal{H}$ intersects only with the boundary of $\mathcal{I}\left(\begin{smallmatrix} D_{11} \\ D_{21} \end{smallmatrix}\right)$, because it contains less than $\operatorname{rank}\left(\begin{smallmatrix} D_{11} \\ D_{21} \end{smallmatrix}\right)$ chemical species. Therefore, only data points that lie on $\mathcal{H}$ can lead to further restrictions of $\mathcal{I}(D_{11})$.
2. $\mathcal{F}$ remains the same, since the spectra corresponding to the boundary of $\mathcal{F}(D_{11})$ are nonnegative and contain at least one zero entry and are therefore also on the boundary of $\mathcal{F}\left(\begin{smallmatrix} D_{11} \\ D_{21} \end{smallmatrix}\right)$. Thus, the intersection of $\mathcal{H}$ and $\mathcal{F}\left(\begin{smallmatrix} D_{11} \\ D_{21} \end{smallmatrix}\right)$ leads to no further constraints regarding $\mathcal{F}(D_{11})$.

The additional information that is inherent in D_{21} can be included with D'_{21} in the same way as for a shared low-dimensional representation in case 1, since $\operatorname{rank}(D_{11}) = \operatorname{rank}\left(\begin{smallmatrix} D_{11} \\ D'_{21} \end{smallmatrix}\right)$. □

It is also possible to only take the ESI of D_{21} to get D'_{21}, the reduced version of this submatrix that contains only the shared components. This is because only the spectra that increase $\mathcal{I}_S$ are additional constraints, and those are a subset of the ESI of D_{21}. The same holds for D_{12} when considering the transposed case.

If the shared chemical components are mixed with others in D_{21}, D'_{21} can be empty. But since there is still information to be gained from this block, another method is needed. It is possible to consider not only the data points, thus the inner polygon, but also the AFS. With this the ambiguity of the shared chemical components is further reduced, compared to the ambiguity in D_{11}. For this purpose the restrictions of the AFS are transformed into the shared low-dimensional representation of the shared block. This is done by using the intersection of the subspace where the data points of D_{11} are located, which we denote by $\mathcal{H}$, and the AFS of $\left(\begin{smallmatrix} D_{11} \\ D_{21} \end{smallmatrix}\right)$. A possible realization for this is a modified ray casting, see [28] for classic ray casting. This way, all information that can be gained from D_{21} and D_{12} of the shared chemical components is included in the shared low-dimensional representation. The influence on the AFS of the non-shared chemical components is also included in this way, so that a further limitation of the ambiguity is possible.

The reduction in ambiguity benefits greatly from the inclusion of the AFS restrictions, but it should be noted that this comes with a higher computational cost and a more complex algorithm.

Another important fact, which is not self-evident, concerns the knowledge of the appearance of chemical components and can be formulated as follows.

Remark 2.15. It is not necessary to know which chemical components are shared, because of the nature of the calculation as described above. Only the number of chemical species underlying each block is important for a correct low-dimensional representation.

Thus, it has been shown how to handle a matrix as in Eq. (1) for both cases. But also more complex patterns or structures of missing values may occur where it is not possible to attribute them to that L-shaped structure. Therefore, the next section looks at some selected structures and how to handle them.

2.6. Application to other block structures with multiple missing blocks

Not only the structure considered in Eq. (1) can be handled when missing values occur, but also more complex ones. First, we look at a structure that looks quite simple. We assume adjacent blocks share at least one chemical component. For the sake of clarity the missing blocks are left empty in the matrix representation.

$$D = \left(\begin{array}{c|c|c} D_1 & D_2 & \\ \hline & D_3 & D_4 \end{array}\right) = \left(\begin{array}{c|c|c} C_1 S_1^T & C_1 S_2^T & \\ \hline & C_2 S_2^T & C_2 S_3^T \end{array}\right) = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \\ S_3 \end{pmatrix}^T$$

Thus D_1 and D_4 are not directly coupled through a shared factor. They still affect each other's ambiguity, even if they do not share a chemical component through D_2 and D_3. This holds because D_1 is responsible for an ambiguity reduction in the shared low-dimensional representation with D_2 as shared block by reducing $\mathcal{F}_S$. The reduced AFS is then transferred to the shared low-dimensional representation with D_3 as the shared block, where D_4 is also included. Therefore, the ambiguity of the chemical components that are not shared with D_4 is also reduced by this inclusion. Instead of a shared low-dimensional representation, this can also be formulated by comparing the resulting feasible bands and recalculating the AFS under these constraints, as described in Fig. 6.

This can be used for even more complex structures like:

$$D = \left(\begin{array}{c|c|c} & & D_5 \\ \hline & D_2 & D_6 \\ \hline D_1 & D_3 & \\ \hline & D_4 & \end{array}\right)$$

The goal is to split D into blocks that have the same shape as shown in Eq. (1), to analyze each one individually and combine the solutions in a further step. A possible solution for decomposing D is

$$\hat{D}_1 = \left(\begin{array}{c|c} & D_5 \\ \hline D_2 & D_6 \end{array}\right), \quad \hat{D}_2 = \left(\begin{array}{c|c} D_1 & D_3 \\ \hline & D_4 \\ \hline & D_2 \end{array}\right), \quad \hat{D}_3 = \left(\begin{array}{c|c} D_2 & D_6 \\ \hline D_3 & \\ \hline D_4 & \end{array}\right).$$

In order to connect all solutions in the end to one, the shared blocks are used (namely D_6, D_3 and D_2). Let us assume that all these shared blocks have the same rank as D. Then, if a feasible solution is found for one block, e.g., $\hat{D}_1$, a solution can be found for all blocks uniquely, since the concentration profiles respectively spectra are shared. This is similar to case 1. However, if not all components of D are present in all shared blocks, then at least an ambiguity reduction can be gained. That means if the solution of one block, e.g., $\hat{D}_1$, is known, the solutions of $\hat{D}_2$ and $\hat{D}_3$ are not necessarily unique, but the corresponding ambiguity can be reduced, since again the components and their ambiguity is coupled through the shared blocks. This approach is similar to case 2.

The main advantage of having multiple blocks is that limiting the ambiguity in one block (e.g., through additional knowledge) results in a limitation of ambiguity for all other blocks.

To further illustrate the theory, some examples are analyzed in the following section.

3. Examples

For each of the cases described in Section 1.1, a simulated data set is used to visualize the results. In addition, two experimental data sets are provided to analyze more complex situations. The data sets are as follows.

Data set 1. *This simulated data set with $k = 100$ spectra and $n = 100$ frequency channels with the rank 3, see Fig. 1, has a missing block whose number of entries is more than 50% of the total data entries. For this data set Theorem 2.13 enables the reconstruction of a representative of the missing block. We follow the procedure described in Section 2.3. The low-dimensional representation is shown in Fig. 4 and the corresponding feasible bands in Fig. 5. The original profiles are shown in Fig. 8 in purple.*

Data set 2. *This is a simulated hyperspectral image (HSI) data set with 50×50 pixels and whose pixel information is written row-wise into the $k = 2500$ rows of the data matrix. Each row represents the spectrum of a single pixel, and each spectrum has $n = 100$ frequency channels, see Fig. 9 for the pure component spectra and concentration maps. The rank of the matrix equals 4. Some pixels have missing spectral values only in the frequency channels 57–100. These pixels with missing spectral information are plotted white in Fig. 10. This white area simulates a sensor saturation. The rank map is shown in Fig. 11. This example relates to case 2, since not all chemical species are present in the first 56 frequency channels and in the pixels that are not affected by incompleteness.*

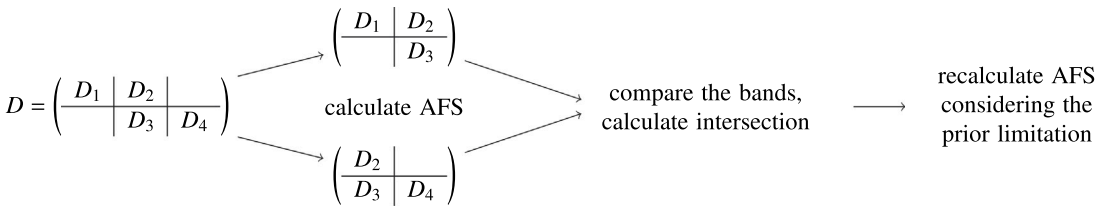


Fig. 6. Schematic representation of dealing with complex block structures.

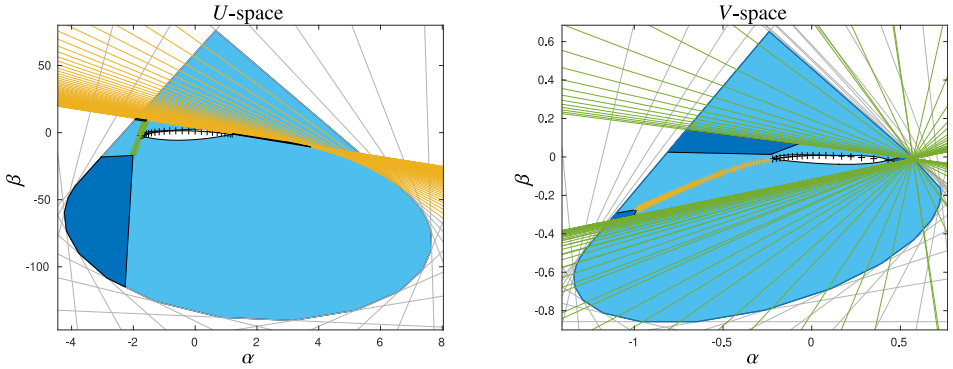


Fig. 7. Low-dimensional representation of the data set 1. The AFS of the shared block is shown in light blue (this is a very large connected set between the inner and outer polygons) and the reduced AFS in dark blue (three isolated much smaller sets). Dual data points and constraints are marked with the same color (yellow and green, respectively). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Data set 3. *An experimental UV/Vis-spectroelectrochemical (SEC) data set on different oxidation states of a phenazine derivative. The SEC experiment records $k = 1518$ spectra each at $n = 1141$ frequency channels. The electrolyte is acetonitrile with 0.1 M of tetrabutylammonium hexaflourophosphate as conducting salt. The working electrode is a gold mesh and the counter electrode is made of a platinum wire. At the beginning the phenazine derivative is present with 0.66 mM. During the measurement the potential is cycled between −0.4 and 1.2 V, with a scan rate of 0.5 mV/s. The equilibrium potentials are 0.029 V and 0.266 V. For chemical reasons, three chemical species are expected. The experimental data set contains regions with missing values due to sensor saturation, making the data set ideal for an application of the methods presented here.*

Data set 4. *This experimental HSI data set including four chemical species is based on a 60×60 grid of pixels. The pixel-wise spectra are stored in the $k = 3600$ rows of the data matrix. Each spectrum comes with $n = 253$ spectral channels. See [36] for details. This data set describes an oil-in-water emulsion, where a region with more or less three chemical species (a subsystem) can be identified using local rank information, see [36]. The selected region is shown in Fig. 19 in the red rectangle. Due to a small signal-to-noise ratio, the complete data matrix has been approximated by its rank-4 truncated SVD. This leads to a higher signal-to-noise ratio. Additionally, we apply a filter approximation to the three-species subsystem in order to cancel any noise or traces of the fourth chemical species in this region. These steps guarantee that the pre-processed experimental data has a sufficient quality to make our approach applicable.*

In the following, these data sets are analyzed using the proposed methods. We present applications to more complex block matrix structures and to subsystems.

3.1. Analysis of case 1 — reconstruction of missing data

Data set 1 can be analyzed in two different ways. The first and simplest approach is to reconstruct a representative of the missing block using Theorem 2.13 and then to analyze the reconstructed data set, for example by using the FACPACK software. This leads to the bands of feasible solutions as shown in Fig. 8. Another way is to use the shared low-dimensional representation described in Section 2.3. Then data points corresponding to the frequency channels and spectra are added to the AFS representation of the shared block, see Fig. 4 in yellow and green. According to duality properties, new data points in the U-space correspond to lines in the V-space and vice versa. This is visualized in Fig. 7 where the dual geometric objects are marked in green and yellow. The impact on the ambiguity is also shown. The AFS of the shared block is shown in light blue and the reduced AFS in dark blue, taking into account the effect of the new spectra and frequency channels. A significant reduction of the ambiguity can be observed. Some parts of the AFS are almost line-shaped or even point-like sets.

The resulting profiles are the same as when the missing block is reconstructed by the first approach. The feasible profiles are shown in Fig. 8.

3.2. Analysis of case 2

Next we treat the more complex case that $\operatorname{rank}(D) > \operatorname{rank}(D_{11})$. Then the missing data cannot be fully reconstructed by adding information from incomplete rows and columns to the AFS representation of the shared block. In addition, there is no shared low-dimensional AFS representation of all species. For example, there is no knowledge of the second part of the concentration profile of a chemical species, if this species contributes only to the frequency range represented by D_{12}. This information is part of the missing block D_{22}. Thus, the information is not available, but there is still information in the incomplete rows and columns (namely in D_{21} and D_{12}) that can be used to reduce the factor ambiguity.

In order to analyze an example problem, we consider the HSI model data set 2. Fig. 9 shows the concentration maps of the four pure chemical components together with their spectra. We assume that the third chemical species is critical (e.g., due to sensor saturation) so that the data of the last 44 frequency channels with the index range 57–100 cannot be used. All rows of D which correspond to such pixels have missing entries in the columns 57–100. The missing value pattern is shown in Figs. 10 and 11. In Fig. 10 this pixel region is shown in white on the concentration maps of all four chemical species. It has considerable overlap with the regions where the other three chemical

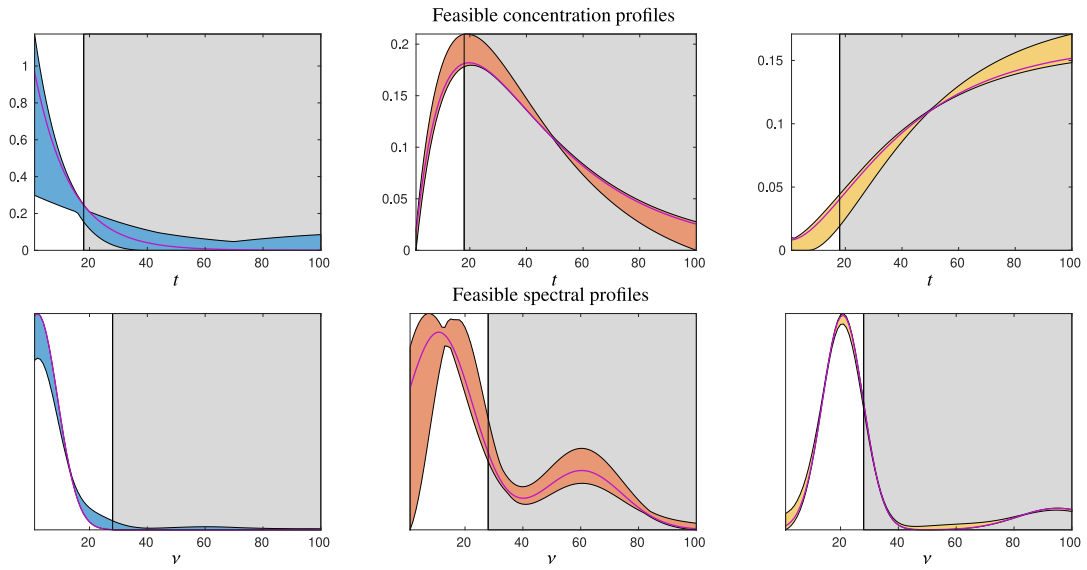


Fig. 8. Bands of feasible concentration profiles and spectra for the data set 1. The original profiles are shown in purple. The reconstructed profiles as derived by the shared block approach are underlaid in gray. The leftmost parts (underlaid in white) of the concentration and spectral profiles correspond to information gained from the shared block D_{11} shown in Fig. 7. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

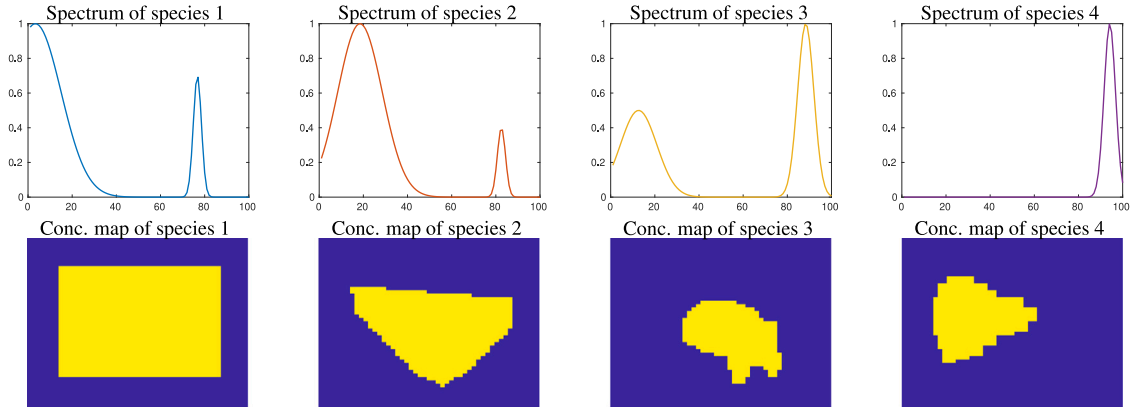


Fig. 9. The HSI model data set 2 with four chemical species. The top row shows the pure component spectra and the bottom row shows the associated concentration maps. Yellow pixels indicate the presence of the species, all with fixed and equal concentration values. Blue pixels indicate the absence of the species on the underlying support layer. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

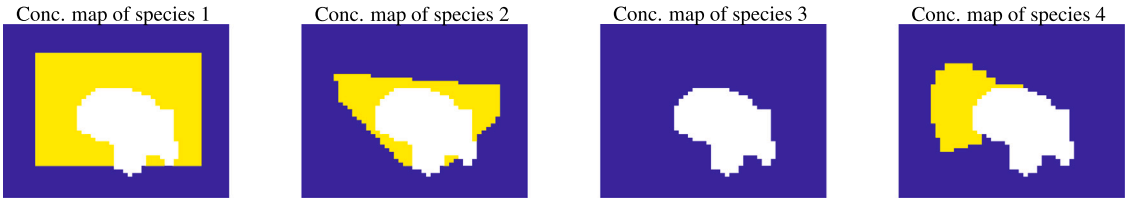


Fig. 10. The presence of the third chemical species (third plot) is assumed to result in sensor saturation only for the frequency channels 57 up to 100. Hence, all pixels to which this species contributes are taken as missing matrix entries in this frequency range. This pixel region is marked in white in the third plot. In other words, each white pixel corresponds to a row of the spectral matrix that contains usable matrix entries only in the frequency channels 1 up to 56. The non-usable pixels (in the frequency range 57 till 100) are also marked in white in the concentration maps of the three other chemical species. There is a considerable overlap with the regions of occurrence of the other three chemical species. In the same way as in Fig. 9 the color yellow indicates the presence and the color blue indicates the absence of the respective chemical species. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

species are found. This model data set meets the requirements of case 2 because the range of missing values (namely the last 44 frequency channels) completely covers the spectral peak of the 4th species, which is only present in the last 20 frequency channels. Therefore, it is clear that the 4th species cannot be reconstructed in the affected pixel region. The only block that contains information about the affected area is D_{21}, where the 4th species does not contribute.

First, the factor ambiguity analysis is based on the shared block D_{11}. Only two chemical species contribute to D_{11} with the rank 2. The low-dimensional AFS analysis results in feasible regions in terms of intervals (in purple) in the U- and V-spaces, see Fig. 12. The associated one-dimensional AFS, which is called a Lawton–Sylvestre plot, for the submatrix D_{11} of D is an $(s_1 - 1)$-dimensional object, in this case two intervals on a line drawn in purple, as shown in Fig. 13. This figure

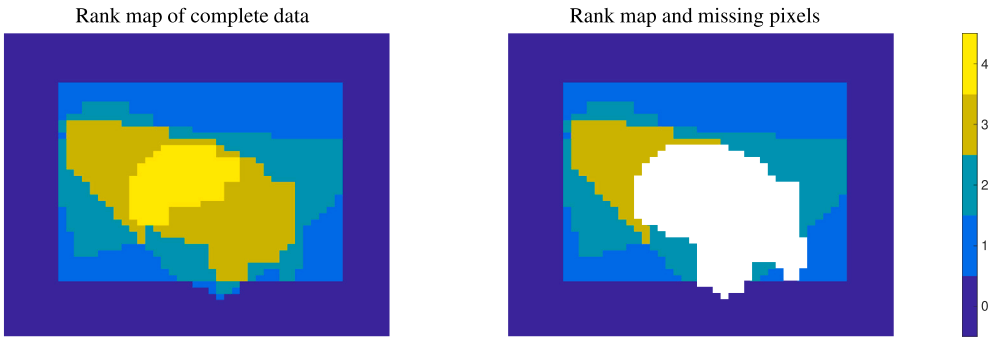


Fig. 11. Left: Rank map over all frequency channels for the complete data set. Possible values are 1, 2, 3, 4. Right: The same rank map, but the pixels which are saturated in the frequency channels 57–100 are marked in white.

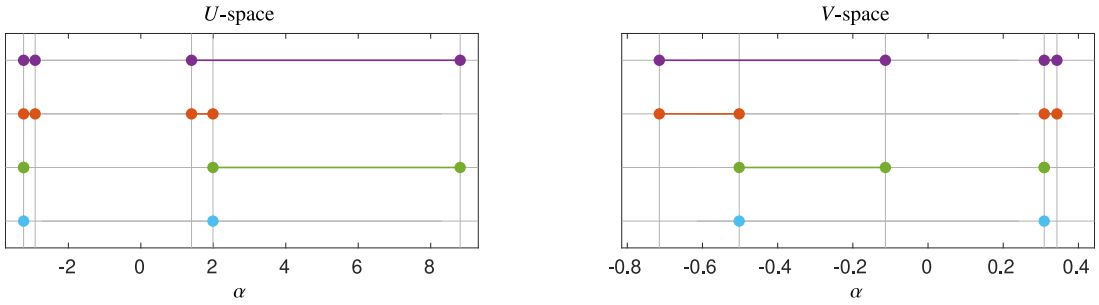


Fig. 12. A Lawton–Sylvestre plot (this is a 1D AFS plot) of the factor ambiguity underlying the shared block D_{11} of the HSI model data set 2 is shown in the top line (left in the U-space and right in the V-space). The purple intervals represent the feasible regions of the two chemical species. Adding the information from D_{21} and D_{12} drastically reduces the ambiguity. First, the red intervals represent the factor ambiguity after adding the information from D_{21}. Second, if we alternatively add the information from D_{12} the purple intervals are reduced to the green intervals. Third, combining the restrictions from D_{11}, D_{21} and D_{12} means to consider the intersection of all these intervals. The intersection consists of only four points, which represent four unique profiles, namely the two spectral profiles and the two corresponding concentration profiles of this two-species system. These points are plotted by circles in cyan. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

illustrates how the additional information of the 2D AFS of the matrix $\begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}$ supports to reduce the ambiguity underlying the shared block D_{11}. Similar steps are shown in Fig. 14 for (D_{11}, D_{12}).

In both cases, the inner polygon corresponding to D_{11} increases, which means for this one-dimensional AFS that the inner endpoints move away from the origin. At the same time, the duality also affects the outer polygons, which means for the given one-dimensional AFS that the outer interval endpoints move closer to the origin. Figs. 12 till 14 illustrate the stepwise reduction of the factor ambiguity. All steps are explained in the respective figure captions. The result is that including the two blocks D_{12} and D_{21} in the factor ambiguity analysis leads to unique spectral profiles and unique concentration profiles for the two chemical species underlying D_{11}.

Furthermore, these two unique factors lead to uniqueness for the remaining two chemical species. Thus the complete system has successfully been analyzed. There is a unique pure component factorization. The four unique spectral profiles are shown in Fig. 15 and the associated concentration maps are plotted in Fig. 16. The missing data is still responsible for an unknown part of the spectrum of the third chemical species which is marked by the gray rectangle in Fig. 15. Similar relations hold for the concentration map of the fourth chemical species (the white area in Fig. 16). The corresponding information that would resolve the unknown parts is contained in the missing block.

3.3. Application to the experimental UV/Vis data set 3

Next, we consider the UV/Vis SEC data matrix with several blocks of missing values. The first step for such more complex structured matrices is to reshape the data matrix so that it can be decomposed into a minimum number of subsystems, each of which has the shape as in Eq. (1). This introductory step simplifies the analysis. Each subsystem can be treated individually, resulting in a factor ambiguity analysis for the full matrix. Considering case 1 of Section 1.1, the full spectral information can be recovered and a representative of the full data matrix can be reconstructed with minimal effort. For case 2 of Section 1.1, our approach provides an easy way to extract a maximum of information about the ambiguity of all chemical species from the given spectral data matrix.

Such complex structures can occur, e.g., when a minimal information loss is desired so that only a few scattered parts of the data are declared invalid. Sometimes, one may prefer to ignore some parts of the data that cause a complex data structure. However, one must pay for this by an increased factor ambiguity. These relationships are analyzed by means of the data set 3. Sensor saturation occurs in several regions, see Fig. 17. This data set can be reshaped to minimize the procedural effort. For this data set, the reshaping operation is illustrated in Fig. 17. After reshaping, only a single shared block is required to analyze most of the data, namely D_2. Then, the blue and yellow parts of the matrix are taken into account for the factor ambiguity analysis.

A submatrix rank analysis reveals that D_1 and D_2 contain two chemical species and that D_3 contains all three of the species that are present in the data set. The next step is to evaluate the constraints determining the factor ambiguity that refer to the shared block D_2. This step follows the procedure in Section 3.2. Since all species are present in D_3, it is also possible to represent all information in the AFS representation of $\begin{pmatrix} D_2 \\ D_3 \end{pmatrix}$. This is analyzed next.

First, the factor ambiguity of (D_2, D_1) is analyzed as in case 1. The accessible information is represented in the Lawton–Sylvestre plot (which is a 1D AFS) of D_2. The feasible intervals are shown in purple in Fig. 18 (left). This result can be embedded in the AFS plot of the three-species system represented by the 2×1 block matrix $\begin{pmatrix} D_2 \\ D_3 \end{pmatrix}$ which takes into account the spectra that are represented in the V-space of D_2, see Fig. 18 on the right. The yellow line represents the orientation of the feasible intervals of the two-species subsystem D_2 within the

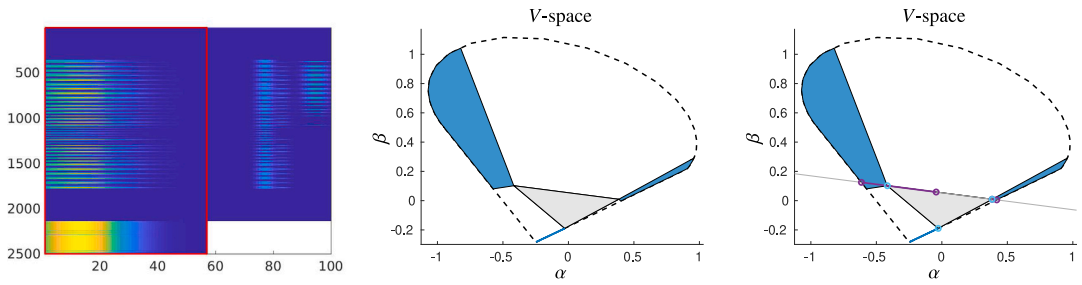


Fig. 13. Left: The mixture data matrix D where the subblock consists of the block matrices D_{11} and D_{21} is surrounded by a red line. Center: The 2D AFS plot of the matrix $\binom{D_{11}}{D_{21}}$ in the V-space. The inner polygon is the gray area and the outer polygon is drawn by a dashed black line. The feasible regions are shown in blue (there is one larger set, one narrow strip and a line-shaped region). Right: The 2D AFS plot is overlaid by the 1D-AFS (a Lawton–Sylvestre plot) of the shared block D_{11} whose rank equals 2. These are the two purple intervals as shown in Fig. 12 (right). The intersection of the purple intervals (namely the embedded Lawton–Sylvestre plot) with the blue feasible regions defines the AFS for the shared chemical species. Thus, using the knowledge of D_{11} has considerably reduced the ambiguity as shown in Fig. 12 by the red intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

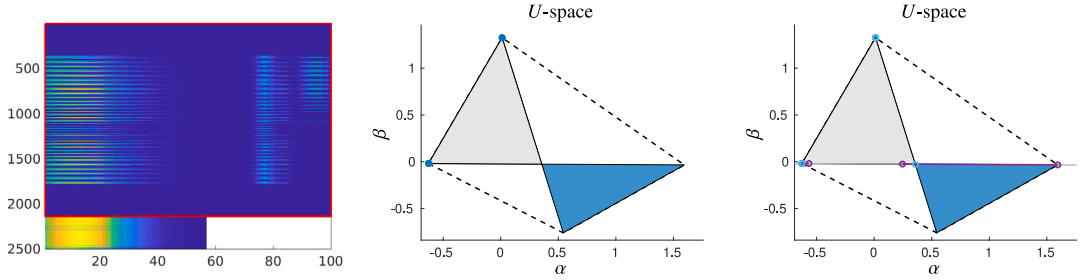


Fig. 14. This is the direct analog of Fig. 13, where the block D_{12} is used to augment D_{11}. Left: The block matrix (D_{11}, D_{12}) is surrounded by a red line. Center: The 2D AFS plot of the matrix (D_{11}, D_{12}) in the U-space. The inner polygon is the gray area and the outer polygon is drawn by a dashed black line. The AFS is formed by one large triangle-shaped set and two point-like sets. Right: Again we consider the intersection with the feasible intervals of D_{11} in purple. The intersection is plotted by green intervals in Fig. 12 (left). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

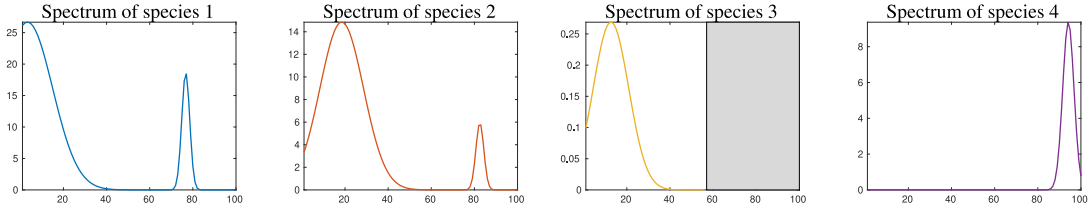


Fig. 15. Reconstructed spectral profiles of the incomplete HSI model data set 2. The missing part of the 3rd spectrum is marked by a gray rectangle. There is not enough information to recover this spectral information.

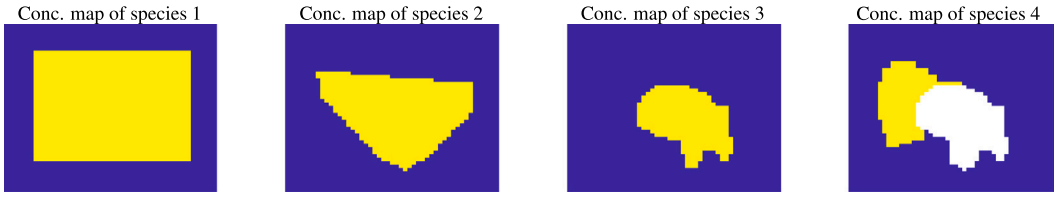


Fig. 16. Corresponding concentration maps to Fig. 15. The white region indicates missing values; there is not enough information to reconstruct the concentration profiles of these pixels. The color yellow indicates the presence and blue the absence of the species. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

AFS plane of the three-species represented by (D_2, D_1). Once again, the correct embedding of the 1D Lawton–Sylvestre plot into the 2D AFS is achieved by determining the expansion coefficients of the profiles of the 1D Lawton–Sylvestre plot within the space of expansion coefficients of the 2D AFS plot.

The intersection of the purple interval (which is partially overlaid by the red interval) and the AFS of $\binom{D_3}{D_2}$ as plotted in light blue results in even shorter feasible intervals for this two-species subsystem. This is shown in Fig. 18 (left) in light blue. The result is that the spectral profiles of two chemical species, which are localized on these feasible intervals, have a considerable impact on the AFS of $\binom{D_3}{D_2}$. The re-calculation of the feasible region for the third species is the much smaller AFS set shown in dark blue in Fig. 18 (right).

Next, we take advantage of the chemical knowledge that the first measured spectrum is a pure component spectrum of a single chemical species. This knowledge allows us to reduce one interval of the 1D AFS to a single point, which is marked in Fig. 18 (left) in red. This additional information on this spectral profile is used to re-calculate

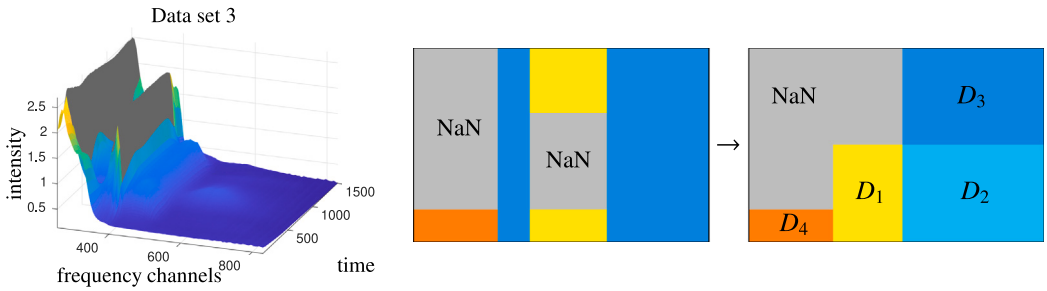


Fig. 17. Left: The experimental UV/Vis SEC data set 3. Gray areas indicate sensor saturation. Right: The original matrix (left) with sensor saturation (these regions are marked by NaN which stands for Not a Number) is rearranged to give the missing data an L-shape (right). This simplifies the factor ambiguity analysis.

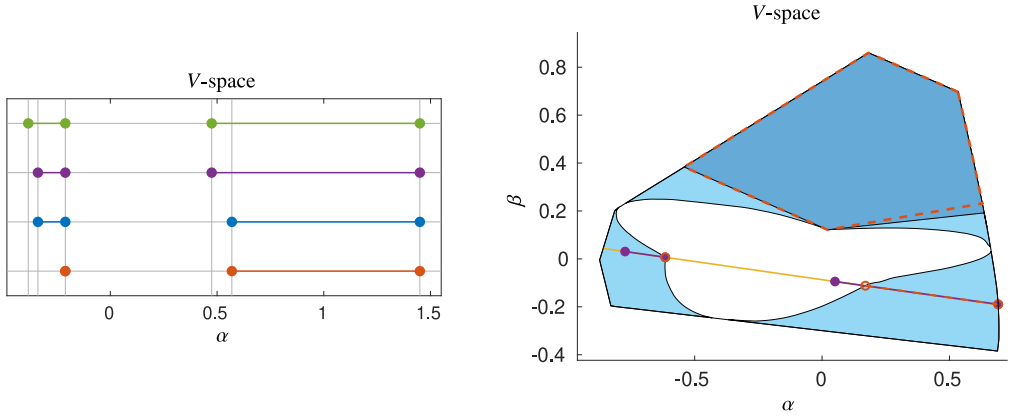


Fig. 18. Left: Lawton–Sylvestre plot (a one-dimensional AFS) of the two-component subsystem associated with D_2 for the data set 3 with the feasible intervals drawn in green. The combined information from D_1 and D_2 results in the somewhat shorter intervals drawn in purple. If additionally, the restrictions through D_3 are taken into account, then the resulting feasible intervals are again somewhat shorter (dark blue). Finally, incorporating the additional knowledge of the spectrum of the first chemical species (this is an additional chemical information by knowledge of the SEC reaction system) results in the red point (the given profile) and a red interval (the still unknown spectrum of the second species). Right: Plot of the AFS of $\binom{D_3}{D_2}$ in light blue. The orientation of the feasible intervals of the two-species subsystem D_2 is drawn by the yellow line, which is overlaid by the feasible intervals of (D_2, D_1) drawn in purple. Taking these two feasible intervals (in purple and also overlaid in red) as constraints for the third chemical species then (the triangle rotation argument) results in the dark blue feasible region for this third species. With the additional knowledge that only the red interval is the true feasible region, the 2D AFS of the third species is reduced. The boundary of this reduced AFS is marked by red dashed lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the 2D AFS. Fig. 18 (right) shows the result where the reduced AFS of the third species is shown in dark blue and with a boundary marked by red dashed lines.

The next step is to include the block D_4. Therefore, a matching shared block has to be selected, e.g., a submatrix of D_1 containing the spectra belonging to the same time coordinates as the spectra of D_4, or a submatrix of (D_1, D_2) again with a selected subset of their spectra. The procedure can then be repeated and the results can be included in the prior shared low-dimensional representation (with D_2 as shared block) by using the resulting feasible concentration profiles.

However, it is not possible to include D_4 in the shared low-dimensional representation due to noise in the proximity of the saturated part, since this would distort the results. It is still possible to approximate the spectrum of the first component in this frequency window, since the first measured spectrum can approximately be taken as the pure component spectrum of the first species.

This data set underlines why our approach is suitable for data with a high signal-to-noise ratio. It also illustrates that knowing a single profile can have a stronger impact on the factor ambiguity than adding a single submatrix (here D_1) to the analysis. See the purple and red feasible intervals in Fig. 18. In the sense of a step-by-step approach it might be useful to check the impact of an additional submatrix such as D_1 in comparison to directly calculating the AFS of $\binom{D_2}{D_3}$ in combination with an application of the knowledge of a certain known pure component spectrum. To check the impact of a submatrix on the ambiguity, the ESI can be calculated and also the resulting AFS in the shared low-dimensional representation can be compared to the AFS of the shared block (in this case the AFS of D_2 can be compared with the AFS of (D_2, D_1)). Therefore, if the effect is sufficiently small, it may not be necessary to consider a submatrix such as D_1 and then to analyze the data set as a complete one.

3.4. Application to the experimental HSI data set 4

In this HSI data set, the focus is on determining pure components and their ambiguity for high-dimensional data by means of subsystem analyses. To this end, we consider a typical HSI data set with a large number of pixels (first dimension of D) and a much smaller number of pure components. This refers to Eq. (2) where the spectral matrix factor is the shared block.

Such an approach is of particular interest when the total system contains more than three chemical species, since increasing the number of components leads to high computational costs for a factor ambiguity analysis (especially when more than three species are considered). Subsystem analyses are much less expensive. Our approach makes it possible to collect all constraints and information about the shared chemical species in the low-dimensional AFS space of the subsystem. This approach saves computation time, but still allows us to represent the information about the shared block in a low-dimensional way. The reduction of the factor ambiguity is achieved in a similar way. If the computational cost is of minor importance, it is also possible to perform the subsystem analysis within the AFS spaces of the full matrix (with its potentially higher number of chemical species) and then to reduce the ambiguity in the same way as for the data set 3.

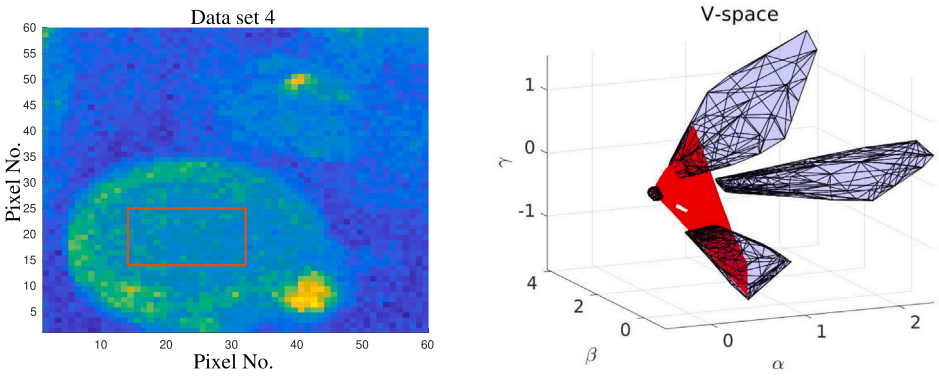


Fig. 19. Left: Mean value image of the data set 4 where the mean values are taken over all frequency channels. The subsystem is marked by a red rectangle. Right: The AFS of the three-component subsystem in red together with the AFS of the four-component system in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For the given data set 4 a subsystem rank analysis yields

$$D = \begin{pmatrix} D_{11} \\ D_{21} \end{pmatrix}, \text{ with } \operatorname{rank}(D_{11}) = s_1 = 3, \operatorname{rank}(D_{21}) = 4 \text{ and } \operatorname{rank}(D) = 4.$$

Hence, the AFS space of this four-species data set D has three dimensions.

As demonstrated above in the analysis of case 2, we consider the intersection of the two-dimensional AFS V-space of the block D_{11} with the three-dimensional AFS space of D. The coordinates of this 2D plane within the 3D space result from computing the expansion coefficients of three different profiles represented in the 2D space with respect to the basis of the singular vectors of the 3D space. The result is shown in Fig. 19 where the 2D AFS set of D_{11} (a connected set with a hole around the origin) is drawn in red and the 3D AFS of D is drawn in blue. The intersection of these two AFS sets is much smaller than the respective 2D and 3D AFS sets. This means that a significant reduction of the factor ambiguity has been achieved. It can be seen that the intersection of both AFS, i.e., all feasible spectra that factorize both D_{11} and three chemical species underlying D, is much smaller than just the respective AFS. This result can be used for the further analysis of the subsystem. For the block D_{11} this means that its underlying factor ambiguity is not given by the relatively large and connected red set, but that it has been reduced to three separated areas. To calculate the reduced AFS it is sufficient to compute the intersection of the AFS of D with the hyperplane spanned by the rows of D_{11}. Without computing the AFS of D, it is possible to determine the intersection in a direct way by a modified ray casting algorithm [28] that uses a point from the hyperplane as a starting point, as well as $s_1 - 1$ non-collinear vectors that define the hyperplane. This makes available all feasible solutions that factorize both D_{11} and the shared chemical components between D_{11} and D_{21}. This step does not require to compute the AFS of D_{11}, but only the hyperplane of the row space of D_{11} is important. Additional information can be obtained by analyzing further subsystems; this will be investigated in future work.

4. Conclusion

The central message of this work is that missing-data MCR analyses can benefit considerably from considering such blocks of data that are ignored when an MCR analysis is applied only to a largest complete data submatrix. This work has demonstrated this for the problem domain of factor ambiguity analysis. We have presented a step-by-step procedure to first compute the AFS sets of the shared block in the U- and the V-space, and then to augment these AFS sets with the spectral information contained in the data blocks D_{12} and D_{21}. These steps may involve data blocks with different numbers of underlying chemical species. AFS sets of different dimensions are then considered. Merging these AFS sets of different submatrices is achieved by embedding the geometric constraints in the AFS space of the highest dimension. The pivotal point of such analyses is the shared data block and its AFS. An application of the presented tool case is not only limited to missing or erroneous data, but is also possible for subsystem analysis and real-time process analysis, where certain parts of the measurements have not yet been recorded, but a prediction of some system properties is desired. The proposed method requires data with a high signal-to-noise ratio, which is a drawback of the present method. This will be investigated in future work.

CRediT authorship contribution statement

Martina Beese: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Tomass Andersons:** Writing – original draft, Conceptualization. **Mathias Sawall:** Supervision, Conceptualization. **Cyril Ruckebusch:** Resources, Methodology, Conceptualization. **Adrián Gómez-Sánchez:** Resources, Methodology, Investigation, Conceptualization. **Robert Francke:** Supervision, Resources, Project administration, Methodology. **Adrian Prudlik:** Resources, Data curation. **Robert Franke:** Supervision, Conceptualization. **Klaus Neymeyr:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to express their gratitude to Dr. Esteban Meija and Phan Huyen Quyen Phung (Leibniz Institute for Catalysis in Rostock, Germany) for providing the data set 3 on the phenazine derivative as well as to Anna de Juan, Department of Analytical Chemistry, Universitat de Barcelona, for providing the HSI data set 4.

References

[1] B. Grung, R. Manne, Missing values in principal component analysis, Chemom. Intell. Lab. Syst. 42 (1) (1998) 125–139.

[2] F. Arteaga, A. Folch-Fortuny, A. Ferrer, Missing data, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemcial and Biochemical Data Analysis, Elsevier, 2020, pp. 615–639.

[3] S. Piqueras, C. Bedia, C. Beleites, C. Krafft, J. Popp, M. Maeder, R. Tauler, A. de Juan, Handling different spatial resolutions in image fusion by multivariate curve resolution-alternating least squares for incomplete image multisets, Anal. Chem. 90 (11) (2018) 6757–6765, PMID: 29697967.

[4] A. Gómez-Sánchez, I. Alburquerque, P. Loza-Álvarez, C. Ruckebusch, A. de Juan, The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values, Chemom. Intell. Lab. Syst. 231 (2022) 104692.

[5] J. Podani, T. Kalapos, B. Barta, D. Schmera, Principal component analysis of incomplete data – A simple solution to an old problem, Ecol. Inform. 61 (2021) 101235.

[6] B. Walczak, D.L. Massart, Dealing with missing data: Part I, Chemom. Intell. Lab. Syst. 58 (1) (2001) 15–27.

[7] B. Walczak, D.L. Massart, Dealing with missing data: Part II, Chemom. Intell. Lab. Syst. 58 (1) (2001) 29–42.

[8] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor, Missing data methods in pca and pls: Score calculations with incomplete observations, Chemom. Intell. Lab. Syst. 35 (1) (1996) 45–65.

[9] S. Wold, C. Albano, W.J. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, M. Sjöström, Multivariate Data Analysis in Chemistry, Springer, Netherlands, Dordrecht, 1984, pp. 17–95.

[10] Y. Beyad, M. Maeder, Multivariate linear regression with missing values, Anal. Chim. Acta 796 (2013) 38–41.

[11] M. Alier, R. Tauler, Multivariate curve resolution of incomplete data multisets, Chemom. Intell. Lab. Syst. 127 (2013) 17–28.

[12] A. de Juan, Chapter 2.5 - Multivariate curve resolution for hyperspectral image analysis, in: J.M. Amigo (Ed.), Hyperspectral Imaging, in: Data Handling in Science and Technology, vol. 32, Elsevier, 2019, pp. 115–150.

[13] M. Alinaghi, R. Rajkó, H. Abdollahi, A systematic study on the effects of multi-set data analysis on the range of feasible solutions, Chemom. Intell. Lab. Syst. 153 (2016) 22–32.

[14] A. Golshan, H. Abdollahi, S. Beyramysoltan, M. Maeder, K. Neymeyr, R. Rajkó, M. Sawall, R. Tauler, A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data, Anal. Chim. Acta 911 (2016) 1–13.

[15] M. Sawall, A. Jürß, H. Schröder, K. Neymeyr, On the analysis and computation of the area of feasible solutions for two-, three- and four-component systems, in: C. Ruckebusch (Ed.), Resolving Spectral Mixtures, in: Data Handling Sci. Technol., vol. 30, Elsevier, Cambridge, 2016, pp. 135–184.

[16] M. Sawall, H. Schröder, D. Meinhardt, K. Neymeyr, On the ambiguity underlying multivariate curve resolution methods, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemcial and Biochemical Data Analysis, Elsevier, 2020, pp. 199–231.

[17] E.R. Malinowski, Factor Analysis Toolbox for Matlab, Applied Chemometrics, Inc., PO Box 100, Sharon, MA 02067, USA.

[18] M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi, M. Maeder, On rotational ambiguity in model-free analyses of multivariate data, J. Chemom. 20 (6–7) (2006) 302–310.

[19] M. Maeder, Y.M. Neuhold, Practical Data Analysis in Chemistry, in: Data Handling Sci. Technol., vol. 26, Elsevier, Amsterdam, 2007.

[20] H. Abdollahi, R. Tauler, Uniqueness and rotation ambiguities in multivariate curve resolution methods, Chemom. Intell. Lab. Syst. 108 (2) (2011) 100–111.

[21] H. Schröder, M. Sawall, C. Kubis, A. Jürß, D. Selent, A. Brächer, A. Börner, R. Franke, K. Neymeyr, Comparative multivariate curve resolution study in the area of feasible solutions, Chemom. Intell. Lab. Syst. 163 (2017) 55–63.

[22] R. Rajkó, K. István, Analytical solution for determining feasible regions of self-modeling curve resolution (SMCR) method based on computational geometry, J. Chemom. 19 (8) (2005) 448–463.

[23] T. Andersons, M. Sawall, K. Neymeyr, Analytical enclosure of the set of solutions of the three-species multivariate curve resolution problem, J. Math. Chem. 60 (2022) 1750–1780.

[24] O.S. Borgen, B.R. Kowalski, An extension of the multivariate component-resolution method to three components, Anal. Chim. Acta 174 (1985) 1–26.

[25] A. Jürß, M. Sawall, K. Neymeyr, On generalized Borgen plots. I: From convex to affine combinations and applications to spectral data, J. Chemom. 29 (7) (2015) 420–433.

[26] M. Sawall, C. Kubis, D. Selent, A. Börner, K. Neymeyr, A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. I: Concepts and applications, J. Chemom. 27 (5) (2013) 106–116.

[27] M. Sawall, K. Neymeyr, A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. II: Theoretical foundation, inverse polygon inflation, and FACPACK implementation, J. Chemom. 28 (5) (2014) 633–644.

[28] M. Sawall, K. Neymeyr, A ray casting method for the computation of the area of feasible solutions for multicomponent systems: Theory, applications and FACPACK-implementation, Anal. Chim. Acta 960 (2017) 40–52.

[29] R.C. Henry, Duality in multivariate receptor models, Chemom. Intell. Lab. Syst. 77 (1–2) (2005) 59–63.

[30] R. Rajkó, Natural duality in minimal constrained self modeling curve resolution, J. Chemom. 20 (3–4) (2006) 164–169.

[31] S. Beyramysoltan, R. Rajkó, H. Abdollahi, Investigation of the equality constraint effect on the reduction of the rotational ambiguity in three-component system using a novel grid search method, Anal. Chim. Acta 791 (2013) 25–35.

[32] M. Sawall, A. Jürß, H. Schröder, K. Neymeyr, Simultaneous construction of dual Borgen plots. I: The case of noise-free data, J. Chemom. 31 (12) (2017) 2954.

[33] H. Minc, Nonnegative Matrices, John Wiley & Sons, New York, 1988.

[34] M. Sawall, C. Ruckebusch, M. Beese, R. Francke, A. Prudlik, K. Neymeyr, An active constraint approach to identify essential spectral information in noisy data, Anal. Chim. Acta 1233 (2022) 340448.

[35] L. Coic, R. Vitale, M. Moreau, D. Rousseau, J. de Morais Goulart, N. Dobigeon, C. Ruckebusch, Assessment of essential information in the Fourier domain to accelerate Raman hyperspectral microimaging, Anal. Chem. 95 (42) (2023) 15497–15504, PMID: 37821082.

[36] A. de Juan, M. Maeder, T. Hancewicz, R. Tauler, Use of local rank-based spatial information for resolution of spectroscopic images, J. Chemom. 22 (5) (2008) 291–298.